

Möglichkeiten und Probleme bei vergleichenden Wortschatz-Untersuchungen zum öffentlichen Sprachgebrauch in der Bundesrepublik und der DDR

Es ist für die Angehörigen einer strapazierten Nation wie der deutschen zweifellos nicht nur unter fachwissenschaftlichem Aspekt von Belang, ob sich aus den nebeneinander bestehenden zwei Verkehrs- und Kommunikationsgemeinschaften auch zwei Sprachgemeinschaften¹ entwickeln oder entwickeln können – mit allen Folgen für das in Regierungserklärungen und vielen offiziellen Reden als eminent wichtig, ja als vielleicht entscheidend bezeichnete, gewiß schon jetzt problematische Bewußtsein von der "Einheit der Nation".²

Aus der DDR häufen sich zur Zeit die Stimmen, die die Existenz zweier deutlich differenter Sprachvarianten (oder gar "Sprachen") schon jetzt behaupten oder doch für die absehbare Zukunft für unausweichlich erklären.³

In der Tat wird die Frage nach der sprachlichen Übereinstimmung – hier verstanden als Frage nach der Übereinstimmung des lexikalischen Inventars und seiner Anwendung – um so wichtiger und dringender, je weniger Hilfen dem Einzelnen (und den kommunizierenden Gruppen) aus gegenseitiger Situationskenntnis oder durch die Anwendung kommunikativer Metatechniken zur Verfügung stehen.⁴ Einerseits ist klar, daß Verstehen und Verständigung nicht nur vom Grad der Wortschatz-Übereinstimmung oder -Differenz abhängen, sondern auch von einer ganzen Reihe außersprachlicher, z.B. psychologischer Faktoren. Klar ist aber auch, daß z.B. Differenzen im Wortschatz den Prozeß des Verstehens sowohl von Texten (in eindirektionaler Kommunikationssituation) wie auch eines Gesprächspartners (in dialogischer Kommunikationssituation) behindern, was wiederum Rückwirkungen auf die Verständigungsbereitschaft haben kann.

Es ist insofern eine legitime und gesellschaftlich notwendige Aufgabe der entsprechenden linguistischen Disziplin, die Frage der Wortschatzdifferenzierung mit aller Sorgfalt und Gründlichkeit so weit wie möglich einer Klärung zuzuführen. Das öffentliche, sowohl publizistische als auch staatliche Interesse zeigt, daß von der Linguistik Antworten erwartet werden, und zwar Antworten, die nicht nur für Linguisten, sondern auch z.B. für die mit dem sprachlichen Ost-West-Thema zwangsläufig befaßten Lehrer

und für andere Interessierte zugänglich und praktisch nutzbar sind.

Natürlich geben wir uns nicht der Illusion hin, an der Wortschatzdifferenzierung selbst etwas ändern zu können. Beeinflußbar können aber ihre Auswirkungen sein, insofern nämlich allgemein zugängliches Wissen über Art, Ausmaß, Schwerpunkte und Tendenzen der Wortschatzdifferenzen die Gefahr von Kommunikationsstörungen oder gar -abbrüchen vermindern und zugleich Hilfen zu ihrer gezielten Überwindung geben kann.

Dies ist eines der – zugegebenermaßen außerlinguistischen – Motive für ein linguistisches Vorhaben, nämlich für ein Projekt, daß das IDS zugunsten seiner Bonner Forschungsstelle für öffentlichen Sprachgebrauch (F. ö.S.) bei der DFG zu beantragen im Begriff ist.⁵ Mit Hilfe dieses Projekts sollen die seit nunmehr 10 Jahren betriebenen Vorarbeiten der F.ö.S. zum "sprachlichen Ost-West-Problem" in erweiterter Form fortgesetzt und zu einem gewissen Abschluß gebracht werden. Sein Titel lautet:

"Ost-West-Wortschatzvergleiche". Vergleichende Untersuchungen zum Wortschatz in Zeitungstexten der Bundesrepublik und der DDR".

Wir denken es uns in 2 Teile zu je etwa 2 Jahren gegliedert; dabei soll der erste Teil dem Ausbau der vorhandenen Textgrundlage und der maschinellen vergleichenden Auswertung dienen, der zweite Teil der Erstellung vergleichender Wörterverzeichnisse, insbesondere der Erstellung eines lemmatisierten Wörterbuchs der ost-west-differenten Lexeme (dazu vgl. unten S. 266f.) sowie der Anfertigung lexikologischer Einzelstudien.

Wir legen Wert auf die Feststellung, daß das Projekt nicht auf die Hervorbringung eines bestimmten Wörterbuchs hin konzipiert ist, sondern auf dreierlei: auf die Bereitstellung von hochgradig variabel abfragbaren Texten, auf die Entwicklung bzw. Weiterentwicklung maschineller und nichtmaschineller Verfahrensweisen für lexikologische Arbeiten und schließlich auf die Hervorbringung einer ganzen Palette von Ergebnissen, von denen eines jenes vergleichende Wörterbuch sein wird.

Wir legen zweitens Wert auf die Feststellung, daß eben dies, die Hervorbringung konkreter Verzeichnisse der Ost-West-Differenzen in benutzerorientierter Form, zu den Zielen mit Priorität gehört, die auf jeden Fall erreicht werden sollen. Und wir legen drittens Wert auf die Feststellung, daß unser Projekt in weiten Teilen quasi als Fallstudie für ein umfassendes textgestütztes Wörterbuchunternehmen im IDS betrachtet werden kann, sofern das IDS eine entsprechende Konzeption – Herr Weinrich hat dies in seinem Vortrag⁶ ja dringend genug empfohlen – entwickelt. Schließlich liegt es auf der Hand, daß ein relativ aufwendiges Projekt wie das hier

skizzierte so weit wie möglich auch Anschlußstellen für interdisziplinäre Weiter- und Folgearbeiten bereitstellen soll. Neben linguistischen Disziplinen wie Terminologieforschung, Textsortenforschung oder angewandte Sprachstatistik möchte ich auf Zeitungswissenschaft und Politikwissenschaft hinweisen. Am wichtigsten freilich erscheint uns der pädagogisch-didaktische Aspekt einer unterrichtsbezogenen Auswertung.

Generell ist von Lehrern verschiedener Schulstufen und -arten häufig darauf hingewiesen worden, daß Themen wie "Sprache und Gesellschaft", "Sprache und Politik" in den Lehrplänen und im Deutschunterricht zwar immer mehr an Gewicht gewinnen⁷, jedoch aus Mangel an verwertbaren wissenschaftlichen Ergebnissen, Verfahrensmodellen und Auswertungshinweisen die didaktische Realisierung oft entweder unterbliebe oder unbefriedigend verlaufe. Es wäre daher sehr zu wünschen, wenn etwa parallel zum Projekt oder sonst im Anschluß daran eine didaktisch orientierte Arbeitsphase der Aufbereitung und Zusammenfassung der wissenschaftlichen Ergebnisse eingerichtet werden könnte, und zwar in enger Kooperation mit entsprechenden Fachleuten.

Es scheint mir nun zunächst notwendig, einige Bemerkungen zu den Voraussetzungen des Projekts voranzustellen, zumal über die Arbeit der Bonner Forschungsstelle vor diesem Kreis seit vielen Jahren nicht mehr berichtet worden ist. Ich nehme in Kauf, daß einiges schon in einem früheren Bericht⁸ erwähnt worden ist.

Die Forschungsstelle für öffentlichen Sprachgebrauch (F.ö.S.) ist – damals noch als "Außenstelle Bonn" – als älteste Arbeitsstelle des IDS im August 1964 auf Initiative von Professor Hugo Moser in Bonn errichtet und dann jahrelang von ihm geleitet worden. Die Gründe für ihre Errichtung lagen zum einen in dem damals lebhaften öffentlichen Interesse an den aus der politischen Teilung resultierenden Folgen für die Entwicklung der deutschen Sprache, zum anderen in der methodisch zumeist unbefriedigenden Lage der mit dieser Thematik befaßten germanistischen Forschung.⁹

Die Aufgabenstellung der F.ö.S. konzentrierte sich daher auf folgende Bereiche:

1. Maschinelle Dokumentation von ost- und westdeutschen Zeitungstexten (als derjenigen Textgattung, in der ein Maximum von Wortschatzdifferenzen aus den verschiedensten Lebensbereichen erwartet werden kann), und zwar in der Form zeitlich gestufter Jahrgangs-Querschnitte;
2. Programmierung und maschinelle Auswertung mit dem Ziel der Erstellung vergleichender Register;

3. Auswertung vorhandener Wörterbücher;
4. Bereitstellung und Auswertung der Sekundärliteratur;
5. wissenschaftlicher Service.

Im Rahmen dieser Bereiche hat die F.ö.S. in den vergangenen zehn Jahren folgende Arbeiten geleistet:

1. Es wurde ein computergestütztes Textkorpus zunächst aus den beiden Tageszeitungen DIE WELT und "Neues Deutschland" (ND) aufgebaut, bestehend aus repräsentativen Querschnitten der Jahrgänge 1954, 1964 und 1969 der beiden Zeitungen (Gesamtumfang z.Zt. ca. 1,8 Mill. lfd. Wörter) sowie einigen weiteren ergänzenden Zeitungstexten. Die Aufnahme entsprechender Querschnitte des Jahrgangs 1974 ist geplant. Ergänzendes Material aus anderen westdeutschen Zeitungen liegt z.T. als Rohmaterial inform von Zeitungssatzstreifen bereit. Jede einzelne Texteinheit (jeder "Artikel") ist weitgehend klassifiziert (u.a. nach genauer Stellenangabe, Autor oder Agentur, Sachgebiet, Intention, Form, Umfang sowie Thema), außerdem sind im Text selbst bestimmte Stellen wie z.B. Überschriften, Eigennamen, Abkürzungen, fremdsprachige Stellen usw. gekennzeichnet.
2. Es wurden Index-, Suchwort- und Textsegmentierungsprogramme entwickelt, mit deren Hilfe ein dv-gestützter systematischer Ost-West-Vergleich unter einer Vielzahl von Vergleichsaspekten möglich wird, außerdem aber – und erstmalig – auch ein zeitlicher Vergleich über jetzt 15, später 25 Jahre west- oder ostdeutscher Wortschatz-entwicklung.
3. Mehrere deutschsprachige Wörterbücher (Duden, Mackensen, Wahrig) wurden insbesondere im Hinblick auf den BRD- bzw. DDR-spezifischen Wortschatz untersucht. Vor allem wird das Ostberliner "Wörterbuch der deutschen Gegenwartssprache" (WGD), das als einziges Wörterbuch systematische Angaben zu Nachkriegs-Neuerungen im Wortschatz und seiner Ost- oder West-Spezifik macht, einer sorgfältigen, später auch maschinell auswertbaren Analyse unterzogen.
4. Die Beobachtung und Auswertung der einschlägigen Sekundärliteratur schlug sich nieder a) im Aufbau einer speziellen Handbibliothek und eines Archivs der Zeitschriftenaufsätze, b) in einer kommentierten Bibliografie der Literatur zum öffentlichen Sprachgebrauch in der BRD und der DDR, die seit Herbst 1974 fertig gesetzt und umbrochen ist, jedoch aus verlagsinternen Gründen lange Zeit auf Eis lag¹⁰, und schließlich c) in einem auch maschinell verfügbaren Verzeichnis der in der Sekundärliteratur behandelten DDR-spezifischen Wörter. Im Rahmen dieses

Teilvorhabens sind bisher über 30 verschiedene materialreiche Arbeiten, meist aus der BRD, exzerpiert und maschinell erfaßt worden. Das Material aus dem WDG und der Sekundärliteratur fassen wir zusammen unter der Bezeichnung "Sekundäres Wortmaterial".

5. In allen Bereichen ihrer Tätigkeit hat die F.ö.S. laufend wissenschaftliche Hilfe bei Forschungsvorhaben innerhalb und vor allem außerhalb des IDS geleistet, und zwar als maschineller Datenservice, als bibliografischer Service und als Arbeitsberatung.

Die Arbeiten der F.ö.S. wurden etwa seit 1969 mit wachsender Menge und Komplexität des Textmaterials und der Auswertungsprogramme sowie mit der Ausweitung auch der sonstigen Aufgaben zunehmend dadurch behindert, daß die verfügbaren Personalstellen in keiner Weise den Arbeitsaufgaben entsprachen. Die Forschungsstelle hat nie mehr als 2 Planstellen für wissenschaftliche Mitarbeiter gehabt; sie hat bis heute weder eine Stelle für Programmierung noch für Datenverwaltung. Die Zahl der Hilfskräfte ging sogar um 40 Prozent zurück. Außer den 2 Planstellen für wissenschaftliche Angestellte sind vorhanden: anderthalb Planstellen für nicht-wissenschaftliche Angestellte (Büro, Datenerfassung), 9 studentische Hilfskräfte (halbtags) einschließlich Raumreinigung, Telefon- und Kopierdienst usw., ferner wird eine Wissenschaftlerstelle provisorisch aus Hilfskraftmitteln finanziert.

Vor allem die viel zu geringe Zahl der Planstellen hatte zur Folge, daß verschiedene Arbeiten wegen häufigen Wechsels oder wegen Kürzung von Hilfskraftmitteln mehrfach unterbrochen oder abgebrochen werden mußten, was insbesondere im Bereich Programmierung und Datenorganisation zu ständigen Störungen und in zwei Fällen zu schweren Rückschlägen führte.^{10a}

Andere dringende Aufgaben wie z.B. statistische und lexikologische Auswertung oder Ansätze zur Sprachverhaltensforschung konnten leider gar nicht erst begonnen werden.

Die eher skeptisch-ironisch gemeinte Vorhersage, die 2 Planstellen, die die Bundesrepublik Deutschland für die Bearbeitung der sprachlichen Differenzierung aufbringe, würden in absehbarer Zeit damit ausgelastet sein, die Arbeiten ihrer DDR-Kollegen zum gleichen Thema zur Kenntnis zu nehmen und auszuwerten, scheint sich jetzt zu bewahrheiten. Denn erstmalig sind 1974 mehr Veröffentlichungen in der DDR zum Thema Sprachdifferenzierung erschienen als in der BRD, nachdem das Verhältnis in den 60er Jahren eher umgekehrt bei 1:3 gelegen hatte.

Wir müssen leider weiterhin¹¹ ein erhebliches Mißverhältnis zwischen öffentlichem, auch staatlichem Zuspruch und tatsächlicher finanzieller Förderung

feststellen. Wenn es nicht gelingt, die Erforschung der Sprachdifferenzierung endlich auch mittelfristig ausreichend abzusichern, ist die Fortsetzung der Fortsetzung der bisherigen Forschungsarbeit mit weiterhin unzulänglichen Mitteln weder aussichtsreich noch praktikabel. Dies gilt mit besonderer Dringlichkeit für den Bereich Textdokumentation.

Es scheint zweckmäßig, einige Erläuterungen zur Bedeutung der Textdokumentation nicht nur für unser Projekt zu geben.

Unter Textdokumentation verstehen wir:

die systematische rechnerorientierte Ermittlung, Erfassung, Bereitstellung und maximal variable Verfügbarmachung eines Textkorpus, das in angebotener Weise für einen bestimmten genau zu definierenden Teilbereich textgewordener Sprache repräsentativ sein soll. Dies ist als Minimalbestimmung zu verstehen; ich bin mir bewußt, daß insbesondere zum Stichwort "maximal variable Verfügbarmachung" noch sehr viel zu ergänzen wäre. Auf jeden Fall enthält diese Minimalbestimmung eine wesentliche Erweiterung gegenüber dem, was man sonst, z.B. im Bundesförderungsprogramm "Information und Dokumentation" des BMFT, unter Dokumentation versteht: hier also nicht nur die Verfügbarmachung von Sekundärdaten ü b e r etwas, z.B. über Texteinheiten, über sprachliche Erscheinungen und dergleichen, sondern schon die Erfassung und Bereitstellung der sprachlichen Primärdaten selbst.¹²

Nur in der Verbindung von primärer und sekundärer Dokumentation hat linguistische Textdokumentation überhaupt auf längere Sicht einen Sinn.

Allerdings ist projektbezogene Textdokumentation in der Regel schon aus Zeitgründen und aus Gründen zwangsläufig zu kurz gesehener Ökonomie eben nicht auf maximal variable, sondern auf spezifisch themen- und zweckorientierte Verfügbarmachung der sprachlichen Daten angelegt; sie tendiert dann allzu leicht in Richtung "Haus-Korpus".

Dies gilt, wenn auch mit deutlichen Unterschieden, eigentlich für alle derzeit vorhandenen Textkorpora zur deutschen Gegenwartssprache, unseres eingeschlossen. Solange es aber keine einheitliche Konzeption, keinen überzeugenden Systementwurf für Textdokumentation zur deutschen Gegenwartssprache gibt, und solange es keine zentrale Stelle gibt, die in der Lage wäre, eine solche Systemkonzeption zu realisieren, die vorhandenen Korpora so weit wie möglich in eine Datenbank zu integrieren, diese systematisch zu erweitern und sie wiederum der gesamten Linguistik in maximal variabler Form zur Verfügung zu stellen, so lange sind wir in Versuchung, immer wieder neue "Haus-Korpora" mit nur eingeschränkter Brauchbarkeit und Verfügbarkeit für den "Rest" der Linguistik zu konstituieren.

Ebenso lange werden wir aber auch Mühe haben, die Ansicht zu widerlegen, Textdokumentation sei in Anbetracht des erforderlichen Aufwandes nicht vertretbar, sei unrentabel; oder weniger kraß: ein kleines Textkorpus sei ja als Sammlung von Beispielsätzen ganz zweckmäßig, aber im Grunde habe doch jeder Linguist seine Beispielsätze mehr oder weniger im Kopf. Noch schwieriger wird es, wenn man darauf beharrt, nicht bloß ein Textkorpus, sondern ein zeitlich gestuftes Textkorpus zu benötigen, was ja deutlich die Absicht der Wiederholung signalisiert.

Textdokumentation als notwendiger ständiger Basis-Service für die gesamte Linguistik – insbesondere aber für die Lexikologie, die auf Korpora wohl mehr noch als andere Disziplinen angewiesen ist –, diese Auffassung begegnet heute noch meist unbehaglichem Kopfschütteln in den einschlägigen Forschungsförderungsinstitutionen.

Die wichtigsten derzeit maschinell verfügbaren Textkorpora zur deutschen Gegenwartssprache sind:

1. Das Mannheimer Korpus des IDS
2. Das Lunder Korpus von Frau Rosengren
3. Das LIMAS-Korpus Bonn
4. Das Saarbrücker Korpus
5. Das Bonner Korpus des IDS.

Zu 1:

Das Mannheimer Korpus des IDS¹³, 1965 bis 1967 erstellt, hatte den Zweck, eine Textsammlung in erster Linie für die "Untersuchung grammatisch-syntaktischer Erscheinungen"¹⁴, und zwar insbesondere im Hinblick auf das damals anlaufende Projekt "Grundstrukturen der deutschen Sprache", bereitzustellen. Es enthält vor allem Romane, Erzählungen sowie wiss. und populärwiss. Darstellungen zu einzelnen Sachgebieten (ca. 60 %), daneben etwas Trivalliteratur, ferner 5 Hefte aus populärwiss. Zeitschriften (darunter 2 Hefte aus der DDR), ca. 90 Seiten aus FAZ und WELT 1966/67 (nur unsignierte Texte jeweils aus der Titelseite), ca. 175 Seiten BILD-Zeitung in rotierendem Auswahlverfahren aus 7 Monaten 1967. Dazu kommen seit kurzem noch einige Texte der Gattung Gebrauchsliteratur überwiegend aus 1970 - 72.

Umfang: ca. 2,5 Mill. lfd. Wörter.

Sonstige Charakteristika:

- keine Auswahl, sondern 100-Prozent-Aufnahme einzelner Werke (außer bei BILD)

- unsystematische zeitliche Streuung
- Klassifikation der Texteinheiten nur nach Groß-Gattungen
- Überrepräsentation bei den Gattungen “Belletristik” und “Populärwiss. Darstellungen” (bestimmter Sachgebiete)
- grammatisch-syntaktisch teilweise “parallelcodiert”.

Zu 2:

Das Lunder Zeitungskorpus¹⁵ in der von Frau Rosengren geleiteten Arbeitsstelle für moderne deutsche Zeitungssprache hat als einziges eine überwiegend lexikologische Zweckbestimmung gehabt, nämlich die Untersuchung des Wortschatzes westdeutscher Zeitungen (Wortschatzstatistik) mit dem (ursprünglichen) Ziel der Entwicklung verbesserter Deutsch-Lehrbücher.

Zusammensetzung:

Setzstreifen-Texte der Zeitungen WELT und Süddt. Zeitung von Nov. 66 bis Okt. 67, und zwar nur der 5 Sparten Meinung, Politik, Wirtschaft, Kultur (Feuilleton), Vermischtes.

Ausgeschlossen bzw. nicht erfaßt werden:

- a) die Sparten Werbung, Sport, Lokales, ferner alle nicht regelmäßigen Sparten und alle Beilagen;
- b) alle nicht signierten Texte (z.B. Nachrichten und Agenturtexte);
- c) alle nicht redaktionellen Texte (Leserbriefe, Roman, Anzeigen etc.);
- d) alles, was größer als 9 Punkte gesetzt ist (Über-, Unterschriften usw.).

Umfang:

WELT-Texte ca. 2,5 Mill. lfd. Wörter

SZ-Texte ca. 1/2 Mill. lfd. Wörter

Es ist damit das derzeit größte aller deutschsprachigen Textkorpora.

Sonstige Charakteristika:

- hochgradige Aufnahme (WELT ca. 50 %, SZ ca. 100 %) der Gesamttexte bestimmter Sparten und Textkategorien
- keine volle Übereinstimmung Streifentext – Zeitungstext
- keine zeitliche Stufung, aber klare zeitliche Fixierung
- Klassifikation nur nach “Sparten”
- Unvollständigkeit der Texteinheiten (Artikel).

Das Korpus steht nach Äußerungen der Verantwortlichen auch für andere Untersuchungen zur Verfügung.

Zu 3:

Das in der Bonner Forschungsgruppe LIMAS konzipierte und inzwischen zu 3/4 fertiggestellte Korpus¹⁶ sollte in erster Linie als Grundlage für morpho-syntaktische und semantische Textanalyse im Rahmen der LIMAS-Arbeiten zur automatischen Sprachübersetzung dienen.

Zusammensetzung:

In Analogie zum Brown-Korpus der englischen Sprache Auswahl aus 500 verschiedenen Quellen zu je 2.000 Wörtern (zusammenhängender Text) unter Berücksichtigung aller wichtigen Fachgebiete, Gattungen und Erscheinungsformen (incl. Zeitungen), beschränkt auf im Jahre 1970/71 in der BRD erschienene Texte.

Umfang:

Endstand ca. 1 Mill. lfd. Wörter, davon 27 Einzeltexte à 2.000 Wörter aus 7 verschiedenen westdeutschen Tageszeitungen.

Sonstige Charakteristika:

- außerordentlich breite, bewußt dünne und gleichmäßige Streuung über nahezu den gesamten Bereich der Schriftsprache
- eindeutige zeitliche Festlegung auf e i n e n Jahrgang
- für die Untersuchung einer einzelnen Textkategorie oder -klasse wie z.B. "Tageszeitung" wegen zu geringer Materialbelegung nicht geeignet
- Klassifizierung nach Sachgebiet und "Texttyp".

Zu 4:

Das Saarbrücker Korpus¹⁷, in der dortigen Arbeitsstelle für masch. Sprachanalyse unter Leitung von Prof. Eggers vor etwa 15 Jahren als erstes maschinenverfügbares Textkorpus im deutschen Sprachraum erstellt, diente als Grundlage für maschinelle Verfahren zur automatischen Satzanalyse. Es ist streng genommen kein Textkorpus, sondern ein Satzkorpus.

Zusammensetzung:

Zweimal 5.500 Sätze aus "Rowohlts Deutscher Enzyklopädie (rde)" und aus dem politischen und Wirtschaftsteil der FAZ 1959/60 (?) (nur signierte Artikel).

Umfang:

Ca. 200.000 lfd. Wörter.

Sonstige Charakteristika:

- sehr homogen infolge Beschränkung auf nur 2 (verwandte) Textsorten

- sehr klein, daher aussagekräftig zunächst nur für die berücksichtigten Textsorten
- zeitlich festgelegt, aber nicht klar abgegrenzt
- grammatisch-syntaktisch gut erschlossen (Angaben über Klassifikation z.B. nach Thema oder Sachgebiet sind nicht bekannt).

Zu 5:

Das Zeitungskorpus der Bonner Forschungsstelle des IDS¹⁸ ist vorhin schon kurz charakterisiert worden. Es ist von Anfang an konsequent ost-west-parallel zum Zwecke der vergleichenden Untersuchung des Sprachgebrauchs ost- und westdeutscher Tageszeitungen – als wichtigste Repräsentanten öffentlichen Sprachgebrauchs – aufgebaut.

Zusammensetzung (z. Zt.):

Auswahl aus den Jahrgängen 1954, 1964, 1969 der WELT und des "Neuen Deutschland".

Umfang (z. Zt.):

Ca. 1,8 Mill. lfd. Wörter.

Sonstige Charakteristika:

- repräsentative Querschnitte in niedriger Auswahlquote (2,5 %) aus dem gesamten Jahrgang der jeweils aufgenommenen Zeitung
- als einziges zeitlich gestuft
- hochgradig klassifiziert.

Eine E r g ä n z u n g des derzeitigen Bonner Korpus durch Einbeziehung anderer Korpora oder Teile von ihnen ist mit unkontrollierbaren Risiken behaftet und daher n i c h t r a t s a m . Jedoch sind einige Korpora durchaus als V e r g l e i c h s k o r p o r a geeignet, wenn auch in unterschiedlicher Weise:

1. Das LIMAS-Korpus ist – nach Fertigstellung – geeignet als (wenngleich nicht repräsentatives) "Durchschnittskorpus" westdeutscher Schriftsprache eines bestimmten, eng begrenzten Zeitraums, vor dem durch geeignete Vergleichsverfahren die Spezifika der westdt. Texte des Bonner Zeitungskorpus, auch in deren zeitlicher Stufung, abgehoben werden können.
2. Das Mannheimer Korpus eignet sich für einen generellen Vergleich zwischen den Großgattungen Belletristik und populärwiss. Darstellungen gegenüber bestimmten Formen der (westdt.) Zeitungssprache, sofern die zeitliche Einbettung außer Betracht bleiben kann.

3. Das schwedische Korpus, ebenfalls zeitlich klar fixiert, eignet sich mit gewissen Einschränkungen als (außerordentlich großes) Vergleichskorpus zu den westdt. Zeitungstexten der entsprechenden Sparten des Bonner Korpus. Ein solcher Vergleich ist möglich, weil das Bonner Zeitungskorpus ebenfalls (u.a.) nach der Kategorie "Sparte" klassifiziert ist.

Für Auswertungsarbeiten, in denen der Rückgriff auf den vollständigen Originaltext in zitierfähiger Weise notwendig ist, ist es allerdings nicht brauchbar.

Ein Korpus, das für die geschriebene deutsche Sprache der Nachkriegszeit oder nur für einen Teilbereich oder einen bestimmten Zeitabschnitt davon repräsentativ wäre, gibt es nicht. Die Konstituierung eines solchen Korpus auf der Grundlage einer auch die soziologischen Bedingungen der Textproduktion und -rezeption einbeziehenden Theorie ist Ziel eines vom IDS beantragten Projekts "Corpus-Erstellung".¹⁹

Auch das Bonner Korpus beansprucht nicht, für "den öffentlichen Sprachgebrauch" oder auch nur "die Presse" in der BRD oder der DDR repräsentativ zu sein. Repräsentativ (wenn auch nur relativ infolge der kapazitätsbedingt niedrigen Aufnahmequote von 2,4 %) ist nur unsere Auswahl aus dem jeweiligen Gesamtjahrgang.

Das Bonner Projekt verfolgt im Bereich der Textdokumentation vielmehr primär das eingeschränkte Ziel, mit vertretbarem Aufwand aktuelle Texte in zeitlicher Stufung zu dokumentieren, an denen sich die verständigungsrelevanten Sprachdifferenzen (und deren Entwicklung) in möglichst Dichte und Vielfalt feststellen und vergleichend untersuchen lassen; wir gehen dabei von der Annahme aus, solche Texte in Vertretern der Medienuntergattung "seriöse überregionale (und regionale) Tageszeitung" am ehesten zu finden. Die auszuwählenden Zeitungen haben grundsätzlich nur dem Anspruch zu genügen, exemplarische Vertreter der jeweiligen Untergruppen, und zwar in ausgewogener Zusammenstellung ihrer wesentlichsten Merkmale zu sein.

Die Gründe für die Wahl gerade von Tageszeitungen als Grundlage unserer Textdokumentation sowie die Grundsätze und Verfahrensweisen der Textaufnahme und der Klassifikation sind im Forschungsbericht 2 des IDS und – in knapper, aktualisierter Form – im Sammelband "Zum öffentlichen Sprachgebrauch..." dargelegt.²⁰

Die wichtigsten Gesichtspunkte sind zusammengefaßt folgende:

1. Zeitungstexte haben gegenüber anderen Publikationsgattungen (z.B. Belletristik, wissenschaftliche Literatur, Gebrauchs- und Fachliteratur) den Vorteil einer sehr breiten thematischen, inhaltlichen, formalen und intentionalen Streuung; sie sind von vornherein hochgradige "Mischtexte".
2. Zeitungen zeigen eine besondere funktionsbedingte Bereitschaft zur Aktualität auch im Bereich der Wortverwendung; sie spiegeln nicht nur aktuelle Tendenzen der Wortschatzentwicklung (in der Schicht der "Gebrauchssprache"), sondern beeinflussen diese auch (Multiplikatoreffekt).
3. Es kann angenommen werden, daß Zeitungen (insbesondere weitverbreitete Tageszeitungen mit breit gestreutem Leserkreis) durchschnittlich in höherem Grade als andere Textgattungen und auch als der umgangssprachliche Sprachgebrauch etwaige Wortschatzdifferenzen zwischen dem öffentlichen Sprachgebrauch in der BRD und der DDR widerspiegeln werden, und zwar nicht nur im Bereich von Politik, Ideologie und Propaganda, sondern auch im Bereich der nicht-ideologischen materiellen und sozialen Alltagswirklichkeit. Der Vergleich von Tageszeitungen wird also – vereinfacht gesagt – annähernd das Maximum an Differenzen ergeben, das im sprachlichen Ost-West-Verhältnis zu finden ist.

Dabei ist allerdings zu berücksichtigen, daß der Abstand zwischen dem in der Zeitung repräsentierten öffentlichen und dem nicht öffentlichen umgangssprachlichen Sprachgebrauch in der BRD und der DDR vermutlich unterschiedlich groß ist, und zwar infolge der in der DDR entschieden größeren Einheitlichkeit und Verbindlichkeit und damit auch Kontrolliertheit des gesellschaftlichen Normensystems und seiner sprachlichen Äquivalente. Andererseits gilt, daß Öffentlichkeit – und damit gesellschaftlich-politische Normgebundenheit – in verschiedener Hinsicht graduierbar ist. Öffentlicher Sprachgebrauch der Zeitungen und nicht-öffentliche Umgangssprache sind auch in der DDR nicht durch eine Kluft getrennt, vielmehr gilt die Verpflichtung zur Übereinstimmung mit dem einen sanktionierten Normensystem – in situations- und rollenspezifischer Graduierung – für jedes Sprechen außerhalb der vertrauten Kleingruppe. Dies als Einwand gegen das Argument, Analyse des ND reproduziere bloß einen von der Bevölkerung abgelehnten "Parteijargon".

Was die Wahl der Zeitungen betrifft, so stand für die DDR das "Neue Deutschland" als in jeder Beziehung führende und den aktuellen (politischen) Sprach-

gebrauch entscheidend mitprägende Tageszeitung außer Frage. Eine ihm unmittelbar vergleichbare Tageszeitung der BRD gibt es nicht. Die Wahl mußte zwischen den damals als seriös anerkannten Tageszeitungen mit der größten Verbreitung, also FAZ oder WELT, fallen. Als gemäßigte liberal-konservative Zeitung konnte die WELT damals (1965) als kapazitäts- und kostenbedingter Minimalkompromiß gewählt werden. Aufgrund der seit 1969 zunehmend einseitigeren Tendenz und des Rückgangs ihres Marktanteils reicht die WELT heute auch als Minimum nicht mehr aus. Wie das ND, so ist auch die WELT prononcierter Vertreter einer bestimmten politischen Linie. Der Vergleich der Wortschätze zweier "Extreme" kann sinnvoll sein, er ist aber nur dann methodisch ungefährlich, wenn man weiß, inwiefern sich dieser politisch engagierte Wortschatz von Wortschatz weniger politisch engagierter Zeitungen (z.B. Regionalzeitungen) einerseits und vom Wortschatz anders engagierter Zeitungen andererseits unterscheidet.

Das Projekt sieht deshalb die Einbeziehung je einer westdeutschen und ostdeutschen Regionalzeitung vor, und zwar ab Jahrgang 1969. Die Einbeziehung von Regionalzeitungen erst gewährleistet auch die Erfassung von bestimmten Wortschatzbereichen des alltäglichen Lebens, die in überregionalen Zeitungen schwach oder überhaupt nicht vertreten sind. Für die westdeutsche Seite soll das Korpus außerdem ergänzt werden durch die "Frankfurter Rundschau" (F.R.), die als exemplarischer Vertreter der "linken Mitte" betrachtet werden kann. Aus Kapazitätsgründen ist diese Ergänzung, obwohl auch für 1969 erwünscht, nur für den Jahrgang 1974 vorgesehen. Ohnehin macht die bisherige Dokumentation der Jahrgänge 1954-64-69 die Aufstockung durch den Jahrgang 1974 und den Rückgriff auf den frühen Jahrgang 1949 erforderlich, um einen Zeitraum von 25 Jahren deutscher Nachkriegs-Sprachgeschichte überblickbar zu machen. Nicht nur die zeitliche Reichweite, sondern vor allem auch die Aktualität des zuletzt gespeicherten Jahrgangs macht die zeitliche Stufung des Gesamtmaterials interessant. Dies bestätigen auch die uns erreichenden Anfragen. Ein zeitlich gestuftes Korpus zur Gegenwartssprache, dessen letzte Stufe fünf Jahre alt ist, wäre in der Tat von eher archivalischem Wert.

Schon mit Rücksicht auf die sehr hohen Kosten jeder maschinellen Textdokumentation, aber auch mit Rücksicht auf die notwendige und erwünschte Multidisziplinarität textbezogener Forschung mußte ein Aufnahmeverfahren gewählt werden, das möglichst vielseitige Auswertbarkeit zu nicht nur hausinternen und nicht nur linguistischen Zwecken, sondern auch unter den Aspekten bestimmter Nachbardisziplinen (wie Zeitungswissenschaft oder Politikwissenschaft) gewährleistete. Die zu treffende Auswahl aus den Zeitungsjahrgängen mußte also so beschaffen sein, daß sie – unter

möglichster Wahrung der Vollständigkeit und Originaltreue der einzelnen Texteinheiten — die gesamte Breite der in einer bestimmten Zeiteinheit (Jahrgang) produzierten Textmenge "modellfähig" abbildete: alle Monate und Wochentage, alle Sparten und Seiten sind entsprechend ihrem Vorkommen in der Gesamt-Jahrgangsmenge in der Auswahl zu berücksichtigen. Nichts wird aus der Auswahl ausgeschlossen; alle Texteinheiten werden vollständig aufgenommen. Ein entsprechend rotierendes Auswahlverfahren stellt die Repräsentativität der Auswahl sicher, allerdings auf der Basis einer (kapazitätsbedingt) relativ niedrigen Auswahlquote von 2,5 Prozent der Gesamtmenge (Maßstab 1:40).

Die Schreibkonventionen und die Klassifikation nach externen Kategorien gewährleisten relativ geringe Informationsverluste und hohe Variabilität der Auswertung.

Dies relativ aufwendige Verfahren hat allerdings zur Folge, daß wir die in vielen Zeitungsverlagen in der Regel kostenlos anfallenden Setzstreifen (TTS-Streifen) nicht so ohne weiteres verwenden können. Sofern wir Lochstreifentexte verarbeiten, müssen sie in gleicher Weise repräsentativ sein, die gleiche Qualität und den gleichen Klassifikationsgrad erreichen wie die von uns selbst erfaßten Texte. Dies macht umständliche Prozeduren des Vergleichens, Identifizierens, Eliminierens, Ergänzens und Korrigierens erforderlich, die den möglichen Leistungsgewinn drastisch reduzieren, nämlich auf etwa 30 bis 50 Prozent gegenüber dem manuellen Verfahren, und auch dies nur bei Einsatz moderner Datenerfassungsgeräte.

Nach Beendigung der Texterfassungsarbeiten wird ein Zeitungskorpus im Umfang von ca. 4,6 Mill. lfd. Wörtern zur Verfügung stehen, dessen Aufbau und mengenmäßige Gliederung die folgende Graphik zeigt:

- Erläuterung: WE = DIE WELT Berliner Ausgabe, ab 1969: Deutschland-Ausgabe
zusätzlichem Berliner Lokalteil
- ND = Neues Deutschland, Berliner Ausgabe
- RZW = Regionalzeitung West (Wuppertaler General-Anzeiger)
- RZO = Regionalzeitung Ost (Leipziger Volkszeitung oder andere)
- F.R. = Frankfurter Rundschau

Mengenangaben
in 1000 lfd. Wörter

		West	Ost		
		WE 49 180	ND 49 160		
		WE 54 216	ND 54 180		
		WE 64 417	ND 64 208		
		RZW 69 ca. 300	WE 69 600	ND 69 250	RZO 69 ca. 150
F.R. 74 ca. 500	RZW 74 ca. 350	WE 74 ca. 650	ND 74 ca. 280	RZO 74 ca. 160	
ca. 500	ca. 650	ca. 2063	ca. 1078	ca. 310	



Stammkorpus



Erweiterungskorpus 1



Erweiterungskorpus 2

Die Graphik läßt erkennen, welche Vergleichsmöglichkeiten sich im Projekt und natürlich darüber hinaus für weitere Interessenten bieten:

Vorgesehen sind alphabetische Vergleichsregister mit Frequenzen und Berechnung der Frequenzdifferenzierung, und zwar

- a) im Vergleich zwischen zeitlich gestuften Texten (= zeitbezogene Differenzen),
- b) im Vergleich zwischen Ost- und West-Texten (= Ost-West-Differenzen),
- c) im Vergleich zwischen Regional- und Überregional-Zeitungen oder politisch konträren Zeitungen bei Zeitgleichheit (= typenbezogene Differenzen),
- d) im Vergleich zwischen bestimmten Textklassen (z.B. zwischen gleichen oder auch verschiedenen Sachgebieten) (= textklassenspezifische Differenzen), und zwar zeitbezogen oder auch ost-west-bezogen.

Wir glauben, daß insbesondere die Möglichkeiten des Vergleichs verschiedener Zeitstufen eine ganz wesentliche Voraussetzung – und Anregung – für diachronische lexikologische Forschung sein können, – wenn auch zunächst nur in Bezug auf Zeitungstexte. Diachronische lexikologische Forschung findet – außer in Bezug auf einzelne Autoren im Bereich der Literaturwissenschaft – so gut wie nicht statt. Ein wesentlicher Grund – ganz sicher nicht der einzige – ist das Fehlen einer geeigneten, entsprechend gestuften Textbasis. Zur Zeit sind wir, wenn wir von deutscher Gegenwartssprache reden, selten in der Lage, genau zu sagen, wovon wir reden, z.B. wie weit denn die “Gegenwart” nach rückwärts und vorwärts reicht. Gemessen an der vorhandenen empirischen Basis wirken manche Aussagen über den Stand der Gegenwartssprache, oder gar über deren Entwicklung, beeindruckend kühn. Wir möchten auf einem Teilgebiet die Voraussetzungen dafür schaffen, daß sich dies ändern läßt. Dabei beanspruchen wir keineswegs, mit unserem Versuch einer diachronischen Dokumentation von Zeitungssprache, für den es keinerlei Vorbilder gab und gibt, das Problem diachronischer Abbildung von Sprache im Corpus empirisch oder gar theoretisch lösen zu wollen. Auch dies gehört zu den Aufgaben des Mannheimer Projekts, deren Lösung allerdings den Bonner Projektarbeiten nicht mehr zugute kommen kann. Wir haben einen praktischen Versuch systematischer zeitlicher Stufung gemacht, weil er und insofern er für unser Vorhaben notwendig war, und wir hoffen, daß sich auch andere ihn zunutze machen werden.

Zum Aufbau des Projekts nach Arbeitsbereichen

Das Projekt ist in acht Arbeitsbereiche (AB) gegliedert, denen jeweils bestimmte Teilziele zugeordnet sind.²¹

- AB 1 Textdokumentation
Teilziel: Bereitstellung des erforderlichen Textmaterials in bestimmter Menge, Gliederung und Qualität
- AB 2 Sekundär-Dokumentation
Teilziel: Bereitstellung von Korrektiv- und Vergleichsmaterial aus Sekundärliteratur und Wörterbüchern
- AB 3 Textklassifikation
Teilziel: Ermittlung stringenter Klassifikations-Kategorien für Zeitungstexte, Textklassenstatistik, Ermittlung textklassenspezifischer linguistischer Textmerkmale
- AB 4 Programmierung
Teilziel: Erstellung von Programmen für alle Arbeitsstufen des Projekts und für externe Benutzung (Service)
- AB 5 Maschinelle Textzerlegung und Auswertung
Teilziel: Erstellung von vergleichenden alphabetischen Wortformen-Registern mit Frequenzangaben unterschiedlicher Art; Ermittlung wortstatistischer Werte
- AB 6 Lexembearbeitung, Lemmatisierung
Teilziel: Erstellung von Listen der differentiellen Lexeme (Grundformen) mit Verwendungsbeispielen
- AB 7 Wörterbucherstellung
Teilziel: Zusammenfassung der in AB 1 - 6 ermittelten diversen Informationen zu allen relevanten diff. Lexeme nebst Verwendungsbeispielen und Bedeutungsangaben zu "Wortartikeln" in einem "Maschinellen Korpus-Wörterbuch" (MKWb)
- AB 8 Frequenz- und bedeutungsbezogene Einzelstudien
Teilziel: detailliertere, fallstudienähnliche Auswertung der in AB 3, 5 und 6 - 7 gewonnenen Materialien und Erkenntnisse, die sich einer Darstellung in Form von Wörterverzeichnissen entziehen.

Erläuterungen zu den Arbeitsbereichen:

Zu AB 1 "Textdokumentation" ist nur noch nachzutragen, daß die Texte in mehreren, jeweils bedarfspezifischen Versionen und Bereitstellungsmodi (je nach Art der beabsichtigten Weiterverarbeitung) und ggf. auch je nach Art des Rechnertyps bzw. der dort vorhandenen Programme bereitzustellen sind. Beides verlangt eine sorgfältige und systematische Kor-

pus- und Datei-Verwaltung.

Die Arbeitsbereiche 2 bis 4 seien hier übergangen, obwohl AB 2 "Sekundär-Dokumentation" u.a. die Analyse des "Wörterbuchs der deutschen Gegenwartssprache" zum Ziel hat, obwohl AB 3 "Textklassifikation" sich mit der äußerst schwierigen Frage z.B. intentionaler Kategorisierung zu befassen hat, und obwohl AB 4 "Programmierung" von fundamentaler Bedeutung für das Gesamtprojekt ist, da sämtliche Arbeitsbereiche von der ersten bis zur letzten Arbeitsstufe auf Programmierunterstützung angewiesen sind.

Lexikologische Bearbeitungsprobleme treten jedoch erst in AB 5 "Maschinelle Textzerlegung und Auswertung" auf; die Diskussion sei also dort fortgesetzt.

Zu AB 5 "Maschinelle Textzerlegung und Auswertung":

Dieser Arbeitsbereich hat vor allem die Aufgabe, laufende Texte in Wortformen-Register unterschiedlicher Art und Zweckbestimmung zu verwandeln, und zwar als Indices (Wortformenverzeichnisse mit Stellenangaben) wie auch als (Sammel-)Register (ohne Stellenangaben, aber mit relativen und absoluten Frequenzen, alphabetisch oder frequentiell sortiert) oder als sog. Differenzregister (das sind Vergleichsregister mit Angabe von Frequenzdifferenzen).

Dazu gehören unter anderem:

- alphabetische Indices zu allen Jahrgängen;
- alphabetische Register und Sammelregister mit Angaben der absoluten und relativen Frequenzen aus bis zu 8 Jahrgängen (z.B. Sammelregister "WELT-Vokabular" oder "Ost-Vokabular" oder "Vokabular 1954" uns.) und dazu nach Bedarf die entsprechenden Frequenzregister;
- alphabetische Differenz-Register mit Angabe der zeitbezogenen, ost-west-bezogenen und typbezogenen Häufigkeitsdifferenzen;
- Register sortiert nach dem G r a d der Differenzen, jeweils zeitbezogen, ost-west-bezogen und textklassenbezogen (das jeweils differenteste Wort steht am Anfang der Liste).

Aus dem entsprechend aufbereiteten Material sollen ferner diverse statistische Werte (Streuung, Verteilung) zur Feststellung von Schwerpunkten und (zeitlichen) Tendenzen der Wortschatzdifferenzen bzw. -veränderungen ermittelt werden. Dabei können, falls erforderlich und falls ausreichende Maschinenkapazität vorhanden ist, maschinenverfügbare Textkorpora wie das Mannheimer IDS-Korpus oder das LIMAS-Korpus als Ver-

gleich herangezogen werden. Als Ergebnis könnte ein dem Rosengrenschens ähnliches Wortformenlexikon, allerdings sicherlich alphabetisch geordnet und mit differenzierteren statistischen Vergleichswerten, entwickelt werden.

Bei der Auswertung von Wortformen-Register bzw. Indices ist jedoch besondere Vorsicht geboten, wie folgende Stichproben-Untersuchung zeigen kann:

Am Wortformen-Index des Jahrgangs WELT 1964 wurden die ersten 2.500 Wortformen des Alphabets untersucht. Die Strecke reicht von A bis *ansonsten*.

Von diesen 2.500 Wortformen kommen 1.491 nur einmal vor, das sind 59,6 Prozent; 1.009 Wortformen sind mehrfach belegt.

Auf der gleichen Strecke, also wieder bis zum Wort *ansonsten*, führt der Wortformen-Index des "Neuen Deutschland 64" 1.071 Wortformen auf. Das Verhältnis gegenüber der Menge der Wortformen im WELT-Index entspricht in etwa den Erwartungen, da wir aus dem Gesamtjahrgang der WELT (entsprechend seinem größeren Umfang) 2,3 mal so viel Text wie aus dem ND aufgenommen haben.

Von den 1.071 Wortformen im ND kommen 663 nur einmal vor; das sind 62 Prozent. Nur 408 Wortformen sind mehrfach belegt. Die prozentuale Differenz der hapax legomena zwischen ND und WELT überrascht nicht, denn bei wachsender Textlänge (der WELT-Text ist, wie gesagt, 2,3 mal so lang) sinkt der Anteil der hapax legomena. Dagegen überrascht der überaus hohe Anteil der hapax legomena an der Gesamtzahl der Wortformen. Nun muß allerdings berücksichtigt werden, daß die speziellen Besonderheiten von Zeitungstexten seltene Wortformen geradezu provozieren. Es gibt eine große Anzahl von Schreibvarianten durch ad-hoc-Abkürzungen, durch Wechsel zwischen Komposition mit und ohne Bindestrich; daneben die Fülle mehr oder weniger geläufiger Komposita und natürlich die große Zahl auch seltener Flexionsformen. Was die Schreibvarianten durch Abkürzungen betrifft, so treten sie im WELT-Index deutlich häufiger auf als im ND-Index; der Grund liegt vor allem in der Fülle von Kleinanzeigen in den WELT-Texten, die solche Abkürzungen in großer Zahl und immer neuen Varianten enthalten.

Gemeinsam vorhanden unter den Wortformen von A bis *ansonsten* sind nur 555! Das sind 22,2 Prozent im WELT-Alphabet und 52 Prozent im ND-Alphabet. Viele von den 555 sind zudem stark frequenzdifferent.

Unter den eben genannten, von mehreren Störfaktoren beeinflussten Voraussetzungen muß die zunächst erschreckend niedrige Zahl gemeinsamer

Wortformen mit großer Vorsicht interpretiert werden. Man kann aufgrund dieser Stichprobe nicht einfach schließen, daß sich das ND mehr als doppelt so stark an die gemeinsame sprachliche Mitte hält als die WELT; oder anders ausgedrückt: daß die WELT zu drei Viertel ihren höchstspezifischen Wortschatz produziert.

Andererseits sollte diese Stichprobe zumindest davor warnen, leichtfertig zu behaupten, im offiziellen Sprachgebrauch der DDR werde viel mehr von der gemeinsamen Norm abweichender Wortschatz verwendet als in entsprechenden Texten der BRD. Es könnte durchaus sein, daß sich bei näherem Hinsehen das Gegenteil als richtig erweist.

Für das weitere Vorgehen zeigt die Stichprobe aber auch, daß neben oder besser gesagt schon vor der ohnehin erforderlichen Rückführung der flektierten Wortformen auf die Grundformen einiger zusätzlicher Aufwand zur Eliminierung von Schreibvarianten getrieben werden muß. Nur dann sind Wortformen-Register vergleichbar. Derartige Eingriffe dürfen natürlich keinesfalls auf der Textebene selbst und auch nicht auf der Verarbeitungsebene der Indices erfolgen, sondern erst auf der Register-Ebene, auf der unmittelbarer Zugriff zu den Belegstellen nicht mehr gegeben und nötig ist.

Zu AB 6 "Lexembearbeitung, Lemmatisierung":

Hauptaufgabe des AB 6 ist es, aus den Wortformen-Registern mit Hilfe des Sekundären Materials unter Rückgriff auf die Kontexte die differenzverdächtigen Lexeme zu ermitteln und aufgrund geeigneter Vergleichs-Operationen ggf. als relevant zu verifizieren. Dabei sind Grenzwerte für die "Signifikanz" von Frequenzunterschieden und die "Relevanz" von Verwendungsunterschieden zu definieren. Zur Ermittlung der differenzverdächtigten Lexeme insbesondere auch für Wortverbindungen (Wortgruppen-Lexeme) werden unterstützende Programme und Suchprozeduren benötigt, da weder davon ausgegangen werden kann, daß das Sekundärmaterial alle Differenzen der Texte schon erfaßt, noch daß alle Lexem-differenzen der Texte durch signifikante Frequenzunterschiede signalisiert werden. Falls vorherige automatische Lemmatisierung nicht oder nur zum geringeren Teil möglich ist, müssen die als relevant erkannten differenten Lexeme in diesem AB lemmatisiert werden.

Hier liegt eine äußerst wichtige Schaltstelle für den Verlauf des Projekts. Wir wissen, daß an verschiedenen Stellen an Verfahren der automatischen Lemmatisierung gearbeitet wird. Nach unserer Schätzung wird keines dieser Verfahren im Laufe der nächsten 2 Jahre eine vollständige automatische Lemmatisierung unserer Texte leisten können. Sehr wohl aber erwarten wir Teillösungen z.B. im Nominal- und Adjektiv-Bereich. Unsere Hoffnungen richten sich dabei am stärksten auf die DFG-geförderte Saar-

brücker Arbeitsgruppe "Automatische Lemmatisierung".^{21a} Die dort erzielten Fortschritte – und im übrigen auch die Möglichkeit und Bereitschaft zur kooperativen Übertragung dieser Ergebnisse auf unser Projekt – werden mitentscheidend dafür sein, ob die eigentliche Lexemanalyse sich auf schon lemmatisiertes Wortmaterial richten kann oder ob die Lexemanalyse sich auf die reinen Wortformen richten muß und die Lemmatisierung ihr erst folgt. Der lexikologische Gesamtnutzen wäre im ersten Fall offensichtlich weit höher.

Zur Lexemanalyse scheinen einige Erläuterungen wünschenswert. Das möglichst ökonomische Auffinden der differenzverdächtigen Lexeme, und zwar aller, ist eines der methodischen Hauptprobleme. Da ja unmöglich jede Wortform "von Hand", d.h. aufgrund einer menschlichen Analyse auf ihre Relevanz hin überprüft werden kann, müssen maschinelle Suchprozeduren eingesetzt werden. Maschinelle Suchprozeduren haben aber durchweg aufgrund der zunächst meist unvermeidlichen Unschärfe der Beschreibung des Gesuchten den von Krallmann²² so benannten ABNO- oder OBNA-Effekt – d.h. ich erhalte zwar alles Gesuchte, aber nicht nur dies, sondern auch vieles andere, das ich nicht brauche (ABNO = All But Not Only), oder ich erhalte nur Gesuchtes, aber leider nicht alles Gesuchte (OBNA = Only But Not All).

Auf unser Problem angewandt ist der OBNA-Effekt unter allen Umständen zu vermeiden und der ABNO-Effekt zweckmäßigerweise so gering wie möglich zu halten.

Welche heuristischen Prozeduren stehen nun zur Verfügung?

Zunächst der Frequenzvergleich. Abgesehen davon, daß er in den untersten Rangklassen nicht funktionieren kann, weil Differenzen dort zu sehr vom Zufall abhängig sind (Signifikanzschwelle), fördert er sowohl zuviel als auch zuwenig zutage: nicht alle frequenzdifferenten Lexeme sind auch gebrauchsdifferent, und vor allen Dingen gibt es Gebrauchsdifferenzen, die sich frequentiell nicht ausprägen.

Hier springt das "Sekundäre Material der im WDG und/oder der Sekundärliteratur als different erklärten Lexeme ein. Aber auch dort ist Vorsicht am Platze: längst nicht alle differenten Lexeme sind schon behandelt und gebucht, und längst nicht alles Gebuchte kommt in unseren Texten vor. Immerhin: die Schnittmenge aus frequenzdifferenten und "Sekundärem" Material dürfte schon nahe an ein minimiertes ABNO heranreichen. Wir würden, wenn das Sekundäre Material entsprechenden Hinweis gibt, auch u n t e r die statistische Signifikanzschwelle gehen, d.h. auch seltenere Lexeme in den Pool der differenzverdächtigen Lexeme aufnehmen. Trotzdem muß weiterhin mit Lücken gerechnet werden. Und selbst wenn keine

eigentliche Lücke auftreten würde, so ist immer dann eine zusätzliche Bestätigung erwünscht, wenn nur eines der beiden Verfahren Differenzverdacht signalisiert.

Dazu zwei Beispiele, von denen das erste noch einen ausgesprochenen Grenzfall darstellt.

Beispiel 1: Kommission. Aufmerksam geworden auf dieses Wort sind wir vor Jahren durch einen Aufsatz von H.J. Gernentz, Greifswald, der aus der Tatsache, daß das IDS damals eine *Kommission für Sprache im geteilten Deutschland* besaß, schloß, das IDS und seine Bonner Forschungsstelle seien unmittelbar gelenkte Werkzeuge der staatlichen Bonner Revanchepolitik.²³ Wir sind darauf dem Gebrauch des Wortes etwas nachgegangen.

Der Vergleich auf Wortformenbasis, hier der Indices der WELT und des ND, fördert folgendes zutage:

Kommission kommt als Simplex vor in

WELT 54	:	13	in ND 54	:	43
WELT 64	:	<u>18</u>	in ND 64	:	<u>32</u>
zusammen		31			75.

Die Frequenzdifferenz ist zwar groß genug, um Differenzverdacht zu erzeugen. Das WGD zeigt jedoch keine Differenz an, ebenso wenig die Sekundärliteratur, — mit Ausnahme jener Stelle bei Gernetz, die wir hier nicht in Rechnung stellen wollen. Der normale Weg wäre die Gebrauchs-Analyse mittels Suchwort-Kontext-Auszügen. Diesen Weg sind wir gegangen. Dabei zeigte sich eine bemerkenswerte Fülle: Neben den Simplicia treten zahlreiche verschiedene Komposita auf, nämlich.

in WELT	:	13	in ND 54	:	22
in WELT	:	<u>17</u>	in ND 64	:	<u>19</u>
zusammen		30			41.

Die Komposita sind allerdings sehr unterschiedlich stark belegt. Am stärksten *EWG-Kommission* mit 21 Belegen in WELT 64 und *Aktivisten-Kommission* mit 35 Belegen in ND 54, danach folgen *Konfliktkommission* (11 x ND 64) und *Plankommission* (7 x ND 64).

Ich möchte jetzt die Ergebnisse der recht aufwendigen Detailanalyse, mit der 2 studentische Mitarbeiter fast 2 Tage lang beschäftigt waren²⁴, nicht in aller Breite wiedergeben. Nur soviel sei erwähnt:

Fast alle Komposita sind Namen oder stehen für solche. Etwa die Hälfte der Simplicia in der WELT sind ebenfalls verkürzte Nennungen von Kom-

posita (in ND etwas weniger); ein Viertel etwa sind Teil von Präpositionalfügungen, meist ebenfalls mit Namencharakter. Ein letztes Viertel ist unklar; es handelt sich offenbar um ad-hoc-Bezeichnungen für Gremien hauptsächlich mit Untersuchungsfunktionen.

Untersucht wurden die Belege dann im Hinblick darauf, wer die jeweilige Kommission eingesetzt hat (Merkmal: amtlich/nicht amtlich) und welche Funktion sie ausüben (Merkmale: Planung + Kontrolle, Untersuchung/Begutachtung, Beratung). Diese Untersuchung lieferte häufig keine klaren Ergebnisse, weil die maschinellen Kontexte nicht umfangreich genug oder aber der Text selbst nicht expliziert genug waren. Es ließen sich nur Tendenzen feststellen: Es gibt offenbar hüben wie drüben Kommissionen jeglichen Typus. Während in den West-Belegen oft nicht klar erkennbar ist, wem eine Kommission verantwortlich ist (möglicherweise weil ihre Verantwortlichkeit tatsächlich unklar ist), steht in den DDR-Belegen, soweit Kommission im Inland gemeint sind, der amtliche Charakter (Staat, Partei, Gebietskörperschaften, Massenorganisationen) fast immer außer Zweifel. Als Funktion scheinen tatsächlich Planung, Aufsicht, Kontrolle ein Übergewicht zu haben, obwohl auch Funktionen wie Vorbereitung und Untersuchung genannt werden. Die andere Bedeutung von *Kommission* in der kommerziellen Bedeutung "Auftragsverkauf" ist in beiden Textgruppen, wenn auch nur 1 bzw. 3 mal, belegt. Kurz: Das Spektrum der Bedeutungen ist auf beiden Seiten gleich, nur werden im ND die Bedeutungskomponenten "amtliches Gremium" und "mit Planungs- oder Kontrollfunktion ausgestattet" häufiger und unzweifelhafter realisiert als in West-Texten.

Es handelt sich um einen Fall an der Grenze dessen, was wir unter relevanter Gebrauchsdifferenz verstehen. Trotzdem hat eben diese Differenz zu dem schweren Mißverständnis bei Gernentz geführt.

Prozedural fällt hier nun aber der immense Aufwand auf, der getrieben werden mußte, bevor überhaupt klar war, ob er sich lohnte. Hier bietet sich nun eine Prozedur an, die viel schneller und einfacher zu dem gewünschten vorläufigen Ergebnis führt: der Vergleich der Komposita. Wenn, wie hier, nur 3 von den 30 Komposita im Westtext und den 41 Komposita im Osttext gemeinsam sind, alle anderen aber überhaupt nur auf einer Seite vorkommen (wir nennen das "Unikate"), dann besteht der dringende Verdacht, daß auch das Simplex selbst – wenn nicht in allen, so doch in vielen Fällen – unterschiedlich gebraucht wird. Dies ließe sich auch an anderen Beispielen zeigen und mit klarerem Ergebnis als hier. Um die Komposita zu einem Wort zu finden, genügt es aber, das maschinelle Lexikon in unmittelbarer Nähe des gesuchten Simplex zu befragen, und zwar, soweit das gesuchte Wort als Bestimmungswort auftritt, in vorwärts-

läufiger, sonst in rückläufiger Abfrage, also mit einem Minimum an Aufwand, was wiederum erlaubt, j e d e s Substantiv und Adjektiv, auch wenn es sich schon um Komposita handelt, dieser Prozedur zu unterziehen.

Es dürfte sich wahrscheinlich empfehlen, diese Lexikon-Abfrage über die Komposita hinaus auch auf Ableitungen auszudehnen: es liegt nahe zu vermuten, daß, wenn etwa *freiheitlich* von der Frequenz her als differenzverdächtig erkannt worden ist, auch *frei*, *Freiheit*, *Befreiung* und *befreien* in die Untersuchung mit einzubeziehen, wobei der Differenzverdacht wiederum durch den Vergleich der frequentiell hochdifferenten Komposita *Freiheitskampf*, *Befreiungsfront* gestützt würde.

Beispiel 2: Rekonstruktion. Das Wort ist relativ selten in allen Texten, liegt also ohnehin nahe an der Signifikanzgrenze. Die Frequenzdifferenz ist gering. Das sekundäre Material gibt allerdings deutlich Hinweise auf Differenz. Komposita liegen weder in Ost noch West vor, das Verb *rekonstruieren*, ebenfalls selten (je 1 - bzw. 2. Mal), bietet gleichfalls keine Hinweise auf Differenz. Wieder wäre die Untersuchung der Kontexte von Hand der einzige Weg zur Absicherung. Aber es gibt andere Hinweise: sämtliche ND-Belege 1964 und 1969 entstammen (4 - bzw. 5 Mal) dem Sachgebiet Wirtschaft, Untersuchgebiet Industrie. Es handelt sich um die DDR-spezifische Neubedeutung: "Neuorganisation eines Betriebs oder Betriebsablauf" (gelegentlich auch für Wohngebäude oder -gebiete im Sinne von Sanierung, und für Fußballmannschaften).²⁵

In der WELT verteilen sich die insgesamt 4 Belege auf verschiedene Sachgebiete und Themen; es handelt sich um die uns geläufige, in ND 54 ebenfalls noch (einmal) belegte Bedeutung: "Nachvollzug einer Handlung, eines Vorgangs". Um diese Differenz zu erkennen, bedarf es, wie wir sehen, nicht der Kontext-Untersuchung; der Vergleich der Textklassenspezifik tut es auch. Textklassenspezifik wird ermittelt als Differenz zwischen der prozentualen Verteilung der Textmenge auf bestimmte Textklassen (hier: Sachgebiete) und der Verteilung der Wortbelege auf die Textklassen. Die abweichende Verteilung ist die Textklassenspezifik; sie ist als Eintrag im Beschreibungsteil des maschinellen Lexikons leicht abfragbar. Im Lexikon des Osttextes wird diese Textklassenspezifik anders aussehen. Die Differenz zwischen beiden Lexikoneinträgen signalisiert Gebrauchsdifferenz.

In einem Punkt – auch dies soll erwähnt werden – haben wir noch keine Lösung gefunden, nämlich für das Auffinden von unbekannten differenten Wortverbindungen, deren einzelne Glieder nicht different sind.

Nicht also: Wortverbindungen vom Typ *konzertierte Aktion*, *friedliche Koexistenz*, die sicherlich gefunden werden. Wohl aber: *offenes deutsches Gespräch*, *Prager Frühling*. Eine Hilfe könnten wohl auch statistische Re-

lationen einbeziehende Umfeld-Analyse-Programme sein, die, um Suchzeit zu ersparen, über so etwas wie einen KWIC-Index operieren könnten. Der maschinelle Aufwand wäre dennoch beträchtlich, solange er sozusagen blind betrieben wird.

Zu AB 7 "Wörterbuch-Kumulation"

Die in AB 7 zu erledigende Hauptaufgabe besteht zunächst in der zusammenfassenden Organisation der in AB 6 als different bestätigten Lexeme samt allen bis dahin über sie gesammelten weiteren Informationen zu einem "Maschinellen Korpuswörterbuch" (MKWb) sowie in dessen weiterer Anreicherung durch Festlegung von "Bedeutungen" zu jedem Lexem und der Herstellung fester Verweise auf jeweils "typische" Verwendungsbeispiele.

Dieses MKWb ist sowohl ein in formatierten Worteinträgen geordneter "Speicher" für eine Vielzahl teils maschinell, teils manuell ermittelter Informationen zu jedem Lexem, zugleich aber auch ein Arbeitsmittel mit ausgedehnten Rückgriffmöglichkeiten auf Primärdateien (Kontexte) und diverse Zwischendateien (Register). Es ist zunächst **k e i n** Wörterbuch in einem druckbaren Sinne.

In zweiter Linie werden die im MKWb gespeicherten Worteinträge, je nach Art des gewünschten Ergebnisses, abgefragt, umformatiert und in Form spezieller Wörterverzeichnisse — mit oder ohne manuell erstellten Kommentaren — ausgegeben; in einem dritten Schritt werden lexikographisch-redaktionelle Arbeiten erforderlich.

Eines dieser Wörterverzeichnisse wird das alphabetische "**V e r g l e i c h e n d e W ö r t e r b u c h d e r O s t - W e s t - D i f f e r e n z e n i n Z e i t u n g s t e x t e n d e r B R D u n d d e r D D R**" sein, dessen mögliche innere Form in dem am Schluß folgenden Schema beschrieben ist.

Es ist zu betonen, daß alle aus dem AB 7 hervorgehenden Wörterverzeichnisse oder Wörterbücher immer nur Korpus-Wörterbücher sein können, die den Wortschatz oder Teilmengen des Wortschatzes (nebst Informationen dazu) der zugrundeliegenden Texte — und nur diesen — enthalten können. Es ist also im Rahmen des Projekts nicht vorgesehen und auch nicht sinnvoll, andere als in den Texten nachgewiesene Lexeme oder andere als aus den Texten ableitbare Informationen in die Wörterverzeichnisse aufzunehmen. Hier können (und wollen) wir also weder "vollständig" sein noch dem WDG Konkurrenz machen.

Andererseits bietet das MKWb, bei vollständiger Ausnutzung seiner Möglichkeiten, Informationen, die bisher nicht, zumindest nicht in dieser Ver-

knüpfung, zu gewinnen waren, nämlich maximal folgende Angaben zu jedem einzelnen Lexem:

- Angaben über relative und absolute Häufigkeit und Rangklasse in allen 14 Jahrgängen, dazu Sammelfrequenzen und Gesamtfrequenz, ferner Angaben über Streuung und Abweichung vom Mittelwert eines Vergleichskorpus;
- Angaben über Schwerpunkte des Vorkommens, z.B. textklassenspezifische (sachgebietsspezifische, intentionale, textgattungsbedingte) Ballungen;
- Angaben über zeitbezogene Häufigkeiten und Auftretenstendenz (zunehmend, abnehmend, gleichbleibend, schwankend);
- Angaben über Ost-West-Differenz, auch in zeitlicher Verschiebung;
- Angaben über zeitungstyp-bezogene Häufigkeiten und Differenzen (west-interne Wortschatzdifferenzen)

und in allen Fällen: Angaben über den jeweiligen Grad der Differenz.

Ferner soll das MKWb zu jedem Lexem Angaben enthalten über:

- belegte Flexionsformen,
- Wortart,
- Vorkommen als Personennamen oder in Namen,
- Bedeutung(en),
- fremdsprachlichen Einfluß,
- ggf. Gebrauchsbesonderheiten.

Zusätzlich abgefragt werden können:

- Vorkommen in Komposita und Ableitungen (falls nicht selbst Kompositum),
- Vorkommen in festen oder halbfesten Wendungen,
- Verweise auf Vorkommen im Sekundär-Material (Sekundärliteratur und WDG),
- Verweise auf typische Verwendungsbeispiele (Kontexte).

Da nicht alle abfragbaren Informationen, noch dazu in unterschiedlicher Sortierung, gleichzeitig darstellbar oder auch nur erwünscht sind, werden mehrere Wörterverzeichnisse unterschiedlichen Typs und Inhalts erforderlich.

Mit diesem Maschinellen Korpus-Wörterbuch wird zugleich eine (maschinelle) Organisationsform und ein dialogorientiertes Arbeitsverfahren ent-

wickelt und erprobt, das – unter entsprechender Änderung des projektspezifischen Selektionsprinzips – als Fallstudie für ein textgestütztes frequenzbezogenes “Wörterbuch der deutschen Gegenwartssprache” moderner Konzeption dienen kann.

Zu AB 8 “Lexikologische Einzelstudien”:

Bei der Bearbeitung des Textmaterials und der Ermittlung der verschiedenen Lexem-Differenzen werden sich einerseits Fakten, andererseits auch neue Fragestellungen ergeben, die in den vorgesehenen Wörterverzeichnissen oder Forschungsberichten nicht behandelt oder dargestellt werden können, sondern eindringenderer Bearbeitung, ggf. unter Heranziehung zusätzlichlichen Materials außerhalb des Korpus, bedürfen. Auch Einzelstudien zu bestimmten methodischen Fragen können für das Projekt erforderlich werden. Da bei der Planung des Projekts nicht vorauszusehen ist, welche Fakten und Fragestellungen sich im Laufe des Projekts als einer eingehenderen Behandlung bedürftig, aber auch würdig erweisen, kann dieser Arbeitsbereich am wenigsten geplant werden. Gedacht ist u.a. an:

- Wortfeld-Untersuchungen;
- Untersuchungen zum Vokabular bestimmter Sachgebiete oder anderer, z.B. intentional bestimmter Textklassen;
- Untersuchungen zur “Wanderung” (West-Ost-Austausch) und zur “Lebensdauer” bestimmter Wörter im Rahmen der durchschnittlichen Vokabular-Konstanz: “Modewort” und lexikalisierte “Neologismen”;
- Untersuchungen zu Unterschieden und Gemeinsamkeiten im politischen Wortgebrauch westdeutscher Zeitungen (WELT/F.R.);
- Untersuchungen zu den Ost-West-Differenzen speziell im “Alltagswortschatz”;
- Untersuchungen zur Didaktisierung der im Projekt gewonnenen Ergebnisse und Erkenntnisse.

Wieviele und welche dieser Untersuchungen ausgeführt werden, hängt außer von der Dringlichkeit und voraussichtlichen Ergiebigkeit auch von der noch verfügbaren Arbeitskapazität ab.

Es ist hier im übrigen darauf hinzuweisen, daß Einzelstudien wie die genannten nicht nur von Projektmitarbeitern, sondern im Rahmen verschiedener Kooperationsformen auch von interessierten externen Mitarbeitern übernommen werden können. Wir haben weder die Möglichkeit noch die Absicht, die durch das Projekt gebotenen Auswertungsmöglichkeiten auch nur annähernd selbst auszuschöpfen.

Anhang

Zu AB 7: Mögliche Form eines Eintrags im "Vergleichenden Wörterbuch der Ost-West-Differenzen"

Hinweis:

Diese Variante geht davon aus, daß neben den differenten auch die nicht differenten Bedeutungen eines Lexems angegeben und belegt werden. Andere Varianten, auch mit anderer Gliederung und weiteren Angaben, sind denkbar.

1. Lemma
2. Belegte Wortformen mit Angabe des Anteils an der Gesamtbelegung
3. a) auf die wesentlichsten Merkmale beschränkte Sacherläuterung und/oder kurze definitorische Angabe der in den Texten realisierten Bedeutungen bzw. Bedeutungsvarianten mit Hinweisen auf die Ost- oder West-Spezifik
b) ggf. Kommentare zu Gebrauchsbesonderheiten
4. Verwendungsbeispiele, geordnet nach den Angaben in 3., getrennt nach Ost und West, mit Fundstellenangabe
5. Vorkommen in Komposita, Ableitungen und Wortverbindungen getrennt in West Ost
6. ggf. Hinweise auf Behandlung in der wortmonographischen Sekundärliteratur (chiffrierter Quellenhinweis)
7. ggf. Hinweise auf Zusammenhänge mit anderen Lemmata
8. Statistische Angaben zu Vorkommen und Verteilung:

	1	2	3	4	5	6	7	8	9	10
	abs. Frequenz rel.	Gew. Diff.	Rang- Klasse	Vertlg. a. Ztg.	Zeitl. Ten- denz	Verteilung Max/Min	Textkl. Sachg.	Schwerpunkt Intent.	Mtlg. form	Diff- Grad
West										
Ost										
Durchschnittl. Abweichung										

Erläuterungen zu 8. Statistische Angaben:

Spalte 1: absolute und durchschnittliche relative Häufigkeit des Lexems (kumulierte Häufigkeit einschließlich flektierter Formen) aus allen West- bzw. Ost-Jahrgängen

Spalte 2: Gewichtete Differenz: relative Frequenzdifferenz, bezogen auf die Rangklasse

Spalte 3: Rangklasse: Stelle des Lexems in der Sammel-Frequenzliste aller Westtexte bzw. Osttexte

Spalte 4: Verteilung auf Zeitungstyp: Angabe über spezifisch ungleiche Verteilung des Lexems auf bestimmte Zeitungen

Spalte 5 und 6: Zeitliche Verteilung

5: Angabe über Vorkommenstendenz (Grad der Zunahme oder Abnahme oder der Schwankung; sonst gleichbleibend)

6: Angabe der höchsten und niedrigsten Jahrgangsfrequenz und der betreffenden Jahrgänge

Spalte 7 bis 9: Spezifisch ungleiche Verteilung des Lexems auf bestimmte Textklassen

7: Schwerpunkt des Vorkommens in Texten bestimmter Sachgebiete

8: Schwerpunkt des Vorkommens in Texten bestimmter Intentionen

9: Schwerpunkt des Vorkommens in Texten bestimmter Mitteilungsforn

Spalte 10: Grad der Ost-West-Differenz.

Untere Reihe: Gemeinsamer Durchschnitt der Werte aus West- und Osttexten und Abweichung davon.

Beispiel: *Rekonstruktion*

1. *Rekonstruktion, rekonstruieren*

2. *Rekonstruktion* (6), *rekonstruieren* (2), *rekonstruiert* (Part., 5)

3.a) I. etwas Zerstörtes oder nicht mehr Vorhandenes aus den Überresten oder mit Hilfe indirekter Zeugnisse möglichst originalgetreu wieder aufbauen oder nachbilden

II. einen Vorgang mit Hilfe von Anhaltspunkten aus der Erinnerung mit allen Einzelheiten noch einmal darstellen, wiedergeben

III. vorhandene Produktionsmittel, auch nichtindustrielle Anlagen, nach dem modernsten Stand von Wissenschaft und Technik umorganisieren oder erneuern, um einen hohen ökonomischen Nutzeffekt erzielen zu können

b) Tendenz zur Gebrauchsausweitung und -übertragung der Bed. III auf nicht-ökonomische Bereiche

4. West:

Zu Bed. I. keine Belege

Zu Bed. II:

Wenn P. Johnsons Rekonstruktion der Vorgänge zutrifft, so wurde Kennedy um eine Option gebracht, die er sicher zu besitzen glaubte und deren Annahme die Rechnung bedeutend logischer macht, als sie 1961 den meisten erschien. (WELT vom 2.6.64 S. 5)

... Peter Palitzsch hat seine Inszenierung der Stuttgarter Uraufführung fleißig rekonstruiert, leider hier mit schauspielerisch meist schwächeren Kräften als dort... (WELT vom 21.2.64 S. 1)

Zu Bed. III. keine Belege

Ost:

Zu Bed. I. keine Belege

Zu Bed. II:

Rekonstruieren wir in Stichworten die Ergebnisse in der Werner-Seelenbinder-Halle am vergangenen Sonnabend. Die demokratische Sportbewegung organisierte in der schönsten deutschen Sporthalle ein großes Hallensportfest ... (ND vom 10.3.54 S. 6)

Zu Bed. III:

Im letzten halben Jahr 1963 konnten wir durch Rekonstruktion 250000 M Kosten einsparen. Leider gibt es aber keinen materiellen Anreiz für die Unterbietung der Materialverbrauchsnormen. (ND vom 28.1.64 S. 3)

Für die Konsumgüter-Binnenhandel übernehmen wir die komplexe Ausrüstung von Objekten, die Beratung über Einsatzmöglichkeiten der Handelstechnik bei der Rekonstruktion von Großhandelslagern, Betrieben des Einzelhandels, Hotels und Gaststätten... (Anzeige in ND vom 29.3.69 S. 4)

Warschau (ADN-Korr.). Das Hüttenwesen in Gorný Slask wird bis 1985 grundlegend rekonstruiert. Dafür sind 36 Milliarden Zloty vorgesehen... (ND vom 8.4.69 S. 7)

5. West: —

Ost:

Zu Bed. III: Rekonstruktionsgruppen, Rekonstruktionsliste, Rekonstruktionsmaßnahmen

6. Bibliografie 205, 764, 1135

7. ****

8. (Statistische Angaben, hier nicht ausgefüllt)

Anmerkungen

- 1 In Anlehnung an L. Zabrocki, Kommunikative Gemeinschaften und Sprachgemeinschaften. In: Folia Linguistica IV, Bd. 1/2, The Hague 1970, S. 2 - 23.
- 2 Die Bedeutung von Sprache und Kultur für die Einheit der Nation wird in offiziellen Erklärungen der Bundesregierung gelegentlich betont, von der Regierung der DDR aber entschieden bestritten. Vgl. schon W. Ulbricht 1970 (Rede auf der 13. Tagung des ZK der SED, ND vom 16. Juni 1970, S. 4).

- 3 Aus der zahlreichen Zeitschriftenliteratur vgl. als Beispiel G. Lerchner, Zur Spezifik der Gebrauchsweise der deutschen Sprache in der DDR und ihrer gesellschaftlichen Determination. In: Deutsch als Fremdsprache 11, H. 5, 1974, S. 259 - 265. – Vgl. ferner den Sammelband: W. Hartung (Hg.), Sprachliche Kommunikation und Gesellschaft (= Sprache und Gesellschaft, Bd. 1), Berlin (O) 1974.
- 4 Vgl. dazu G. Ungeheuer, Sprache als Informationsträger. In: Sprache – Brücke und Hindernis, München 1972: "Je weniger situative Hilfen aus pragmatischer Erfahrung der Gegenwart oder Vergangenheit zur Verfügung stehen, je mehr das Verständnis einer sprachlichen Formulierung allein abhängt von lexikalischen und grammatikalischen Mitteln der verwendeten Sprache, um so unsicherer und um so länger ist der Prozeß des Herstellens eines Verständnisses" (S. 41). – Zum Zusammenhang von lexikalischer Differenz und Verstehen bzw. Verständigung vgl. auch M.W. Hellmann, Wortschatzdifferenzen und Verständigungsprobleme – Fragen bei der Erforschung der sprachlichen Situation in Ost und West. In: M.W. Hellmann (Hg.), Zum öffentlichen Sprachgebrauch in der Bundesrepublik Deutschland und in der DDR (= Sprache der Gegenwart, Bd. 18), Düsseldorf 1973, S. 142 ff.
- 5 Der Projektantrag ist Mitte August 1975 bei der Deutschen Forschungsgemeinschaft offiziell eingereicht worden.
- 6 H. Weinrich, Die Wahrheit der Wörterbücher, in diesem Band. S.
- 7 Eine Umfrage bei den Kultusministerien im Jahre 1968 zeigte, daß in den Deutsch-Lehrplänen fast aller Bundesländer Themen wie "Sprache und Gesellschaft", "Sprache und Politik" oder gar expressis verbis "Sprache im geteilten Deutschland" (Nordrhein-Westfalen) enthalten sind. (Einzelheiten dazu in dem in Anm. 8 zitierten "Bericht" von M.W. Hellmann, S. 18 f.). Andererseits ist seit Hugo Mosers Schrift "Sprachliche Folgen der politischen Teilung Deutschlands" (1962) keine einzige brauchbare zusammenfassende Einführung mehr erschienen.
- 8 Vgl. M.W. Hellmann, Bericht über die Arbeit der Forschungsstelle Bonn des IdS. In: Zum öffentlichen Sprachgebrauch in der Bundesrepublik Deutschland und in der DDR. [Anm. 4], S. 15 - 34. Der Bericht gibt den Stand von 1971/72 wieder.
- 9 Walther Dieckmann hat in seinem Aufsatz "Kritische Bemerkungen zum sprachlichen Ost-West-Problem" in: Zeitschrift für deutsche Sprache 23, 1967, S. 136 - 165) diese Forschung einer grundsätzlich berechtigten, wenn auch wohl negativ überakzentuierten Kritik unterzogen. Einige seiner Argumente waren auch für die Konzeption der Bonner Forschungsstelle im Jahre 1964 relevant.
- 10 Nachdem durch Druckkosten-Zuschüsse aus öffentlichen Mitteln die Finanzierungsprobleme beseitigt werden konnten und die Bibliographie selbst redaktionell auf den neuesten Stand (Sommer 1975) gebracht worden ist, wird ihr Erscheinen nun für den Winter 1975/76 erwartet.

- 10a Die ohnehin unzureichenden Stellen wurden, vor allem bei studentischen Hilfskräften, Ende 1975 nochmals radikal gekürzt.
- 11 Vgl. schon die 1970 geäußerte Kritik an der "paradoxen Situation...", daß man ein Forschungsgebiet von anerkannt hoher gesellschaftlicher Relevanz und öffentlichem Interesse "zwar einer Forschungsstelle überträgt, diese aber so dimensioniert, daß sie nicht in der Lage ist, die von ihr erkannten Probleme in der notwendigen Interdisziplinarität auch zu bearbeiten" (M.W. Hellmann, Bericht [Anm.8], S. 32.
- 12 Das IdS hat im Jahre 1973/74 die unangenehme Erfahrung gemacht, daß unter den staatlichen Förderungsinstitutionen der BRD ein für Textdokumentation Zuständiger nicht zu finden war.
- 13 Vgl. Forschungsberichte des Instituts für deutsche Sprache, Bd. 3 (1969), besonders die Aufsätze von M.W. Hellmann (S. 25 - 54) und U. Engel (S. 75 - 84). Jetzt zum Teil ersetzt durch K. Bayer, K. Kurbel, B. Epp: Maschinelle Textverarbeitung im Rechenzentrum des IdS (= Forschungsberichte des Instituts für deutsche Sprache, Bd. 14, Tübingen 1974.
- 14 U. Engel, Das Mannheimer Corpus. In: Forschungsberichte des Instituts für deutsche Sprache, Bd. 3, 1969, S. 82.
- 15 Inger Rosengren, Ein Frequenzwörterbuch der deutschen Zeitungssprache. Die Welt - Süddeutsche Zeitung. Lund 1972, insbesondere S. Vf. (Vorwort) und IX bis XXIV. Vgl. auch dies., Ein Frequenzwörterbuch der modernen Zeitungssprache - wie und wozu? In: Beiträge zur Linguistik und Informationsverarbeitung 14, 1967, S. 17 ff; dies., Wort und Wortform, in: Studia Linguistica 23, 1969, S. 103 ff.
- 16 R. Glas, Ein Textkorpus für die deutsche Gegenwartssprache. In: IRAL-Sonderband der GAL 43, Heidelberg 1974, S. 112 - 115.
Vgl. auch R. Glas, Gestalt des LIMAS-Korpus. In: A. Hoppe und Mitarbeiter, Das LIMAS-Korpus. Aufbau sowie Beispiele der Auswertung einer Repräsentation geschriebener Texte der deutschen Gegenwartssprache. Bonn 1975 (Manuskript; Druck in Vorbereitung).
- 17 Hans Eggers und Mitarbeiter, Elektronische Syntaxanalyse der deutschen Gegenwartssprache. Ein Bericht. Tübingen 1969, bes. S. 1 - 19.
Aus den Arbeiten von Hans Eggers und seinen Mitarbeitern hat sich der DFG-Schwerpunkt "Elektronische Sprachanalyse" entwickelt, dessen Teilvorhaben "Automatische Lemmatisierung" für die lexikografische Textanalyse von großer Bedeutung ist.
Vgl. W. Klein und R. Rath, Automatische Lemmatisierung, Teil 1 und 2, (= Linguistische Arbeiten Bd. 10); Saarbrücken 1971.
H. Zimmermann, Das Lexikon in der maschinellen Sprachanalyse, Frankfurt 1972;
R. Dietrich, Automatische Textwörterbücher - Studien zur maschinellen Lemmatisierung verbaler Wortformen des Deutschen (= Linguistische Arbeiten Bd.2), Tübingen 1973;
Aspekte der automatischen Lemmatisierung, Teil 1 und 2, (= Linguistische Arbeiten Bd. 12), Saarbrücken 1972.
- 18 Vgl. Manfred W. Hellmann, Zur Dokumentation und maschinellen Bearbeitung von Zeitungstexten in der Außenstelle Bonn. In: Forschungsberichte

- des Instituts für deutsche Sprache, Bd. 2, Mannheim 1968, S. 39 - 125.
- 19 P. Nikitopoulos/U. Engel, "Corpus" – Theoretisch fundierte Erstellung zweier repräsentativer Corpora für die geschriebene deutsche Sprache a) zwischen 1871 und 1914, b) seit 1949. Bei der DFG eingereicht August 1975.
 - 20 Vgl. oben Anm. 18 und 8.
 - 21 "Arbeitsbereiche" sind als verlaufsunabhängige Sektoren aus dem Gesamt-Arbeitsbereich zu betrachten. Ihnen entsprechend werden Arbeitsgruppen gebildet mit allerdings wechselnder Arbeitskraftbelegung und Schwerpunkt-bildung.
 - 21a Vgl. oben Anm. 17.
 - 22 D. Krallmann, Maschinelle Analyse natürlicher Sprachen, in: Nichtnumerische Informationsverarbeitung, H. 96, 1968, S. 286.
 - 23 Vgl. H.J. Gementz, Zum Problem der Differenzierung der deutschen Sprache in den beiden deutschen Staaten. In: Weimarer Beiträge, Jg. 13, 1967, S. 463 - 468, hier S. 463. Die Stelle ist kurz kommentiert bei M.W. Hellmann, Bericht [Anm. 8], S. 16, Anm. 1.
 - 24 Ich danke Astrid Leymarie und Eckart Rüther für die intensive Mitwirkung.
 - 25 Vgl. die im Anhang genannten Textbeispiele. Der Trainer der DDR-Fußballmannschaft sagte während der Fußball-WM in einem Interview im west-deutschen Fernsehen (Juni 1974), daß "der DDR-Fußball rekonstruiert" werde oder worden sei (Hörbeleg).